DOCUMENT RESUME

ED 131 122 TM 005 848

AUTHOR Roberts, A. O. H.
TITLE Foibles and Fallacies in Educational Evaluation.
PUB DATE Apr 76
NOTE 23p.; Not available in hard copy due to marginal legibility of original document; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Achievement Gains; Analysis of Covariance; *Compensatory Education Programs; Criterion Referenced Tests; Educational Innovation; Elementary Secondary Education; *Evaluation Methods; Federal Programs; Norms; *Program Evaluation; Research Design; *Research Problems; Statistical Bias; Tests of Significance; Test Wiseness

ABSTRACT

Federal assistance for special educational programs makes necessary the regular study of evaluations of thousands of innovations in compensatory education, bilingual education, and reading programs. The results are reported to the President and to Congress. However, investigating organizations find only a few programs with adequate evidence and thousands with faulty evaluation designs. Some of the most common faults are discussed, with examples. There are other factors which lower hopes. If greater numbers with real evidence could be found, knowledge would increase even without an increase in the number of exemplary programs. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED131122

TM005 848 BEST COPY AVAILABLE

Foibles and Fallacies in Educational Evaluation

A. O. H. Roberts

RMC Research Corporation

April 1976

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Introduction

Since the beginning of Federal assistance to compensatory education, the recipients of funds have taken part in annual studies of the success of their innovations; ostensibly to weed out the unsuccessful ones, to improve some, and to demonstrate the worth of the more successful with a view to dissemination. Also, one would hope, to justify the continuation of public funds. Reports go to Congress and to the President from the U.S. Office of Education and from the National Institute of Education. In turn USOE and NIE fund professional organizations to undertake studies of evaluations of compensatory (Title I), bilingual (Title VII), and reading programs (Title VIII). These professional organizations cast their nets as wide as they can and draw responses from many hundreds of programs in receipt of funds under one or more of these Titles; and part of each response is, or should be, an evaluation. About one third of these evaluations are written by subcontractors or consultants, and the rest by specialists in school district offices. By far the greater proportion of these evaluations are valueless as demonstrations; notice that it is not that they prove the program to be without value but that they provide no useful evidence one way or the other. It has been this way from the beginning, and only in part because the difficulties are great; but it is astonishing to see how pervasive, persistent and elementary are some of the avoidable wrong practices. Hawkridge, Chalupsky and Roberts (1968) examined over 1,000 program evaluations, selected 98 for site visits and found only 21 that met the criteria. Tallmadge (1974) screened about 2,000 looking for exemplary projects and found 136 of which he subsequently had to reject all but six. Bowers, Campeau and Roberts (1974) searching for exemplary reading programs, could find only 26 out of an initial 1,520. These evaluations by consultants, research organizations, and school district specialists represent a great deal of work and much of it expensive and wasteful; worse still it robs at least some programs of the opportunity for recognition. Lastly, wading fruitlessly through literally thousands of pages in search of usable evidence is at least frustrating for those who evaluate these evaluations. What follows should be seen as a constructive attempt to increase the proportion of useful evaluations from programs.

There are two main headings in what follows. In the first are discussions

of some frequently found but avoidable errors, and in the second, some examples of difficulties usually beyond the control of educators. Perhaps there should have been a third, of practical difficulties and of sources of bias whose explicit recognition is demanded, but for which the only actions are reasonable allowance or approximate corrections, or at least discussion. Here would be found errors resulting from regression to the mean, from volunteering, from attrition or from loss of data.

Major Fallacious Approaches in the Evaluations

Very few programs pass even the minimal requirements. This section describes some of the inaccurate and inappropriate approaches that are found, not just occasionally, but with considerable frequency. Those described here appear so often that they can be considered to be serious trends in a field where the background of varying approaches and the resulting controversy has made sound evaluation more essential than ever.

Use of Criterion-Referenced Tests

Scores of evaluation reports appear with "criterion-referenced testing" as the central, if not the only theme; with no comparison group, no attempt to give meaning to the figures quoted and often only the scantiest description of objectives. In this form it is difficult to see how testing can argue success, although it is easy to understand its popularity; for several years now writers have been actively promoting the belief that if innovations do not live up to expectations, it can only be because tests measured what was not being taught, did not measure what was, did it the wrong way and with score conversions which were inappropriate. For these criticisms, criterion-referenced testing is the "perfect" answer--

- It specifically and intentionally precludes the very comparisons that norm-referenced testing makes possible.
- Those with a vested interest in the success of the innovation have the unchallengeable control of objectives and curriculum, and even of the definitions used; and these need have no relationships to existing systems.
- They alone determine what constitutes successful achievement or "mastery" of objectives, frequently in terms of quite

arbitrarily chosen percentages. (We have seen these ranging from 60 percent to 95 percent, even within a single study.)

- There can be no checks on the validity or reliability of their tests, or on objectivity or consistency of scoring.
- Disappointing results can be blamed on teachers, definitions of the instructional objectives, or standards set too high.
- They can even give purely subjective standards a cloak of respectability by converting results to tallies.

Given the same freedoms to make the rules of the game and name blank cards, one need never lose at Poker or Bridge.

Here is an actual example which though somewhat extreme is not completely atypical. The innovators of a program laid down their own objectives (not given) and provided their own evaluative tools "to assess accomplishment for each of the designated subject areas" (no examples or descriptions). They dismissed the state-mandated standardized tests as being unsuitable, and although these were given to students, they made no use of the data collected. An objective was said to be "mastered" if 70 percent or more of the students completed it. The same instruments were used for pretest and for posttest, after which differences in proportions mastering the objectives were tested with a correlated means t-test; this gave them two bites at the same cherry, since if "mastery" was not attained (below 70 percent managing the objective) then significant "gains" could perhaps be shown. For example, the six pupils in a grade 1 improved their mean proportion of correct responses from 0.04 on the pretest for social studies to 0.09 on the posttest with a t value of 3.21, significant at better than the 1 percent level. In passing, this program managed to produce some of the highest t-values ever seen--one, a value of 36.37, achieved using a sample of only 12; another of 27.67 from a sample of 28; and what must surely be an all-time record, 52.17 on 53 pupils. (They modestly put this as significant at better than the 1 percent level.)

When only 67 percent of the objectives were completed, they "concluded that the instrument and objectives must be reevaluated for grade level appropriateness and content validity." In fact, a fair number fell below the 70 percent level, but this was always laid at the door of design of instrument or choice of objective--never the result of failure of treatment.

This is perhaps a kinder explanation than that in another case we encountered, where an evaluator who was also the innovator blamed the poor showing on unnecessarily strict standards set by teachers in deciding whether the objectives had been achieved. He promised to reeducate those teachers for the next year!

Although the aims are not new, the stress being put on specific naming of objectives, on counting pupils who attain each, and on checking pupils, objectives, and teaching methods wherever testing shows failure, are profitable uses of criterion-referenced testing. Used properly, this approach can raise signals at any one or more points where attention is called for; it may be that the objective needs to be reconsidered or defined, or that the standard of judging is inappropriate, or that the treatment needs modification. That in itself points clearly to the limitations of criterion-referenced testing--at least as it is being used: It is flexible enough at every level to be considered plastic: each school or district chooses its own objectives, which may have little to do with the basic intent of the funding, and may even be at variance with it on occasion; if too few achieve an objective their numbers can be increased by changing the standards, and we never encountered any attempt to demonstrate the objectivity, validity, or reliability of the scoring system. This statement appeared in one report: "Student participants performed better on criterion-referenced tests than they did on standardized or commercially prepared instruments. This performance lends encouragement to the continued development of instruments for (such) education."

The strength of a chain is the strength of its weakest link--and that is no less true for a chain of logic. The use of these head-counts (percentage of pupils passing objectives) for significance--or confidence testing with correlated means t-tests--gives no confidence whatever, if for example, one has no test of the objectivity of teachers' judgments, or of the appropriateness of the specification for the criterion percentage.

If now, to find corrections for these objections--

- universal objectives are set (i.e., the same objectives are set for all schools);
- the scoring system is made objective;
- the standards set are tied to typical performances instead of being arbitrary (as, for example, "70 percent will pass 80 percent of the objectives")

--then comparisons become possible, but the test is then norm-referenced and

becomes subject to the criticisms to which we referred earlier.

Tests of Significance of Gains

Especially as an attempt to plug the holes left by the fashionable criterion-referenced testing, one of the most ubiquitous experimental designs being used to show the "worth" of educational innovation is the derivation and statistical testing of gain scores (i.e., posttest minus pretest). The statistical test applied is usually the correlated means t-test, although sometimes the Wilcoxon Signed Ranks Test is used. When the gain scores result from criterion-referenced tests, we can usually dismiss the demonstration. But even results from recognized achievement tests cause problems more often than not. When these scores are in the form of grade equivalents, it is simply an exercise in futility and a red herring, at best; depending upon the grade level, a gain of a mere three to five months over a year, for a sample of 17 pupils would be likely to be significant at the 0.1 percent level or better--and even that could be just maturation, or practice effect, or both, and independent of even trivial education. But when the scores being used are standard scores, or worse still (as we frequently found) are the original raw scores, it is not even possible to translate the gains into meaningful terms at all.

In general, programs are obviously planned to achieve specified goals, and not to produce research findings. Evaluation is secondary, and in practice a significant proportion of programs subcontract this aspect out to individual consultants or to one of several specialist organizations; roughly one third of all programs have their evaluations done in this way. When this is done late in the process, these consultants and organizations find themselves in the role of "hired guns" defending the program's claims and funding, but in terrain not of their choosing. If called upon to do so in time, these evaluators could in many cases plan to collect more useful data; but when simply given raw data already collected, they will be under some pressure to present summaries in the most favorable light. The remedy of course lies in making greater efforts to ensure that minimum standards for evaluation are met before funding the innovation.

One organization in particular is being employed to do evaluations by school districts from all over the nation with innovations in reading, bilingual education, and other areas. In general, this organization does good work, but it has made a fetish of significance testing of gain scores,

devoting a sizable proportion of its research effort to it. This produces large numbers of "significances" which no doubt please those innovators without statistical sophistication, but often cloud some other, and more dubious, results. In one case, the report contained no less than 204 correlated means t-tests, mostly of raw scores. Predictably, more than half of these were significant at the 0.1 percent level; only 41 were "not significant" in spite of the fact that 60 of the samples contained 14 or fewer students. Every class was tested separately and again as part of the grade level. One sample of 17 produced a t of over 25--for which the evaluator, with a modesty which belied his zeal, claimed a significance of "better than 0.1 percent"; in fact, it had the astronomical value of well beyond one-in-a-googol (i.e., one-in-10^{100})! Almost exactly half of this evaluation report of over 200 pages was devoted to this type of reasoning, and most of that in tabulation and bar graphs.

One does not need to be a mathematical statistician or an educational philosopher to see the fallacy of this approach, but its implications and impacts should be considered carefully. The rough mathematical equation following should be regarded just as a foundation for reasoning.

$$t = \frac{\sqrt{(\text{Sample size} - 1)} \times (\text{Difference between pre- and posttest scores})}{\sqrt{(\text{Sum of variances of the two tests}) - (\text{Twice the geometric mean of the variances} \times \text{correlation of the two tests})}}$$

or, as a very close approximation, but much more concisely

$$t = \frac{\text{Gain}}{\text{Average Standard Deviation}} \times \sqrt{\frac{\text{Sample size} - 1}{2(1 - \text{Correlation})}}$$

Now, it should be easy to see that if only the correlation changes, the value t is smallest when there is no correlation at all between pre- and posttest scores;* in which case just what is the rationale for subtracting the one from the other? Subtracting horses from sheep?

On the other hand, however, if the two tests are thoroughly reliable, and are measuring just the dimension of interest, the correlation will be

* A colleague points out that negative correlations would make t still smaller. Has anyone seen a negative pretest-posttest correlation lately?

high; and as it gets closer to unity, the t value goes up like a balloon. For example, if the correlation between the tests is 0.84 (not a dramatic figure) and there were 33 pupils, we will be multiplying the ratio of gain-to-average standard deviation by a factor of 10. Thus, using grade equivalents for illustration (for which a fairly typical average standard deviation is about eight to ten months), a gain of four months over a whole year would produce a t value of between 4 and 5--significant at around the 0.1 percent one-tailed level.

Now, particularly in the lower grades, a combination of a year's maturation, practice effect of the first testing, with disadvantaged pupils in an ordinary traditional classroom has typically been increasing vocabulary, reading skills, and even basic mathematics by a grade equivalent growth of about 7 months. For a typical class of 25 pupils, and a pretest-posttest correlation of .87 (an actual figure) we would get a t value of about 6.8, with a one-tailed significance of better than the 0.000025 percent level! To interpret this as an indication of dramatic success would be ridiculous.

Few educational innovators are also trained researchers; in fact, only the largest districts have departments that can deal with statistical analysis. For many teachers who already are inclined to point to "the happiness of the children" as a demonstration of success, this kind of statistical glitter is misrepresentation which they will find difficult to resist. For those who are evaluating the report, it is clutter with a nuisance value at best, and otherwise a source of additional computational demands in an attempt to derive meaningful information. For example, one can guess at the correlation between the two tests, then multiply the t value, if given, (or the estimated value, if not) by $\sqrt{\frac{2(1-r)}{N-1}}$. The result is an estimate of $\frac{\text{Gain}}{\text{Average SD}}$, which is a measure of important educational change <u>if</u> the tests are standardized.

The crucial flaw in this approach is <u>not</u>, of course, that there is anything wrong with the statistical procedure itself or even that it does not test the hypothesis proposed; it is that the hypothesis is a trivial one, not worthy of testing. What exactly is the null hypothesis implied? It is that:

"No increase of learning has occurred over the period."

Some learning is taking place even in the absence of teaching; maturation, what children learn from one another, and even what they learned of test-taking itself can all be expected to create change. The correct null

hypothesis then should be:

> "Change in educational conditions has brought no change in the rate of increase of learning."

This hypothesis is tested when, with appropriate care,

- a control or comparison group is used; or
- comparison is made with the rate of increase in the same group before change in educational conditions, or
- comparison is made with increases in classes previous to educational change; or
- some reasonable basis exists for establishing an expectation of increase in the absence of educational change.

The phrase "with appropriate care" above is vital.

There are occasions when the hypothesis, "No change has occurred," is appropriate, although it is probably safe to say that this is never so when considering educational increase. The exceptions are when no change can reasonably be expected without intervention. For example, it is reasonable to expect no significant change in affective measures unless changes have been made in the environment. Examples are self-concept, and teacher and parent attitudes. It is noteworthy that t-tests in these areas are frequently non-significant. The same organization mentioned earlier repeatedly did such tests; here are some figures from the three separate evaluations (different districts) in which affective measures occurred.

- Of 26 educational changes, only one was not significant; but of five measures of affective change, all five failed.
- Of 66 educational changes, only four were not significant; but of 75 affective measures, 44 failed the test.
- Of 90 educational changes, 18 were not significant; but of 30 affective tests, 20 were non-significant.

Misuse of Analysis of Covariance

A more sophisticated version of t-testing is analysis of variance together with its offshoot, analysis of covariance. It is found much less frequently since it demands more expertise. Nevertheless, it has less value than some of its protagonists would believe, being for the most part a means of measuring confidence in observed change, rather than dimension of change.

When applied to pretest and posttest scores, the same limitations apply as for testing of gains. Analysis of covariance, in particular, is occasionally found misused.

This procedure is sometimes used to make "adjustments" for starting differences between treatment and comparison groups on pretests. Theoretically, this makes it possible to compare gains of dissimilar groups. When these starting differences are themselves non-significant, such adjustments do little harm, have at least a superficial logic, but serve little purpose. But when the differences are large, adjustments cannot be supported, most especially when the control group has the lower pretest performance.

Analysis of variance was the precursor and foundation for analysis of covariance. It was devised by Sir Ronald A. Fisher early in this century. He worked primarily in areas of anthropology and agriculture, where measurements were almost invariably the most refined type--ratio measurement. For this type there is a true zero, with considerable confidence that equal intervals at widely separated points can be compared; such measures are lengths, weights, volumes, values, and counts. Educational measures are <u>not</u> of this type; they are sometimes termed interval measures (the next lower type in terms of information provided) but even this may be deceptive, since it implies equality of intervals. There is no direct way that the equality of these intervals can even be tested; and only the most tenuous way in which rough equivalence can be inferred (through assumption of normal distribution, for example). Adjustment through analysis of covariance extrapolates the scale for the lower group upwards, and that for the superior group downwards. Even from two widely separated regions of scoring from a single test, but still more so from different levels of tests, there can be no assurance that the two groups are being measured on the same scale, or even for that matter on precisely the same continuum. There are other more cogent objections to this process, but they call for lengthy statistical argument. Those who can benefit from more sophisticated discussions of the problems involved should refer to the articles by Huck and McLean (1975), and Porter and Chibucos (1975). We will rest our case, but in agreement with Tallmadge and Horst (1974)[1], we should reject conclusions based only upon such evidence.

This should not be seen as a criticism of the analytic process, but of

[1]Tallmadge, G. K., & Horst, D. A procedural guide for validating achievement gains in educational projects. (Revised) Los Altos, California: RMC Research Corporation, December 1974.

one use to which it is frequently put. In general, though, the indications deriving from analysis of variance seem more useful as starting checks than as arguments for success. They certainly should not be used for major sculpturing of unsuitable data. The Encyclopaedia Britannica has a section dealing with the work of Sir R. A. Fisher in general and analysis of variance in particular. It concludes with "Unfortunately the finest statistical treatment will not compensate for poorly selected units of observation or measurement."

Practice Effect

An issue which has received rather scant and cursory attention, but which has the potential for a major upset of many reports on special educational programs, is that of the effects of practice alone on retest scores. In the literature, there are some brief caveats about the increases to be expected from "test sophistication" or from "test interactions" but with few estimates of the size of the effect or its duration. See, for example, Campbell and Stanley in Handbook of Research on Teaching (N. L. Gage, Ed., 1963, p. 175).

In the field, there is a strong tendency for innovators to discount the possibility of bias as a result of practice effect--except when it can be turned to profit; or when there is a fear that the bias can tell against the innovation. The arguments usually given for ignoring the threat are:

- A time interval of seven months or more between testings (the most frequent lapse of time) is enough to wipe out gains resulting from familiarity with tests or with individual items.
- Use of a parallel test nullifies practice effect.
- The effects are allowed for, when a control group is used.
- Modern children are so familiar with test situations that they have already reached a plateau where further testing can add nothing more.
- The effects are small enough to be ignored.
- For practical purposes there will be no further increase from practice alone, from the third testing onward; or, another form of the same argument--

- By intensive coaching spurious benefits can be exhausted before the first testing, so that gains are uncontaminated.

How valid are these arguments? The last sounds plausible enough, except when capital is to be made of the absolute measure of achievement from such normed conversions as grade equivalents, percentiles, stanines, or standard scores; then we would have to remember that the norming sample did not receive the benefits of such coaching. Even this is not enough. Bright students can benefit more than duller ones, from such advice as: "Eliminate one or more obviously wrong alternatives and then, if still in doubt, deliberately guess from amongst the remaining choices." As a result, subgroups would be identified for which the treatment appeared to be a success.

But the most cogent counter-argument is this: Here, as so often happens when dealing with living things, biases tend to be positively correlated and not random; and therefore additive, not cancelling. Who are the ones aimed at by the various compensatory education acts? They are those

- from poorer environments
- in ill-equipped schools
- with severe reading disabilities
- with more dependence upon drill and less upon understanding
- from non-English speaking homes, particularly immigrants
- who are very young, and at the threshold of their educational experience.

And who are the ones with the least contact with tests, with least familiarity with test-taking conditions and with the content of tests, and most subject to the debilitating effects of anxiety and failure? Who have most to gain from experience, from practice and specific instruction? Precisely the same categories of students.

On a priori grounds we would expect there to be effects, at least for some varieties of tests, more especially in cognitive areas. Over the years these expectations have been dealt with under such headings as "Practice Effect," "Test Sophistication," "Test Interaction," and (as a result of Huff's book (1961) Score--The Strategy of Taking Tests) "Test Wiseness"--a concept which Millman, Bishop and Ebel (1965) discussed in some depth and which in

turn sparked attention from several others about that time. But although there has been a great deal of lip-service and some experimentation, both reasoning and demonstrations have been honored more in the breach than in the observance--in recent years on a very large scale indeed. If repeated testing alone can make significant increases in later scores, then a substantial proportion of compensatory education demonstrations of all sorts (Title I, Title VII, and Title VIII included) for the last ten years fall under a large cloud of suspicion; perhaps some failures of demonstrations can even be traced to this as will be seen. By far the greater proportion of all evaluations of compensatory education innovations depend entirely upon a show of gains over two or more testings, with many such gains being statistically significant but practically minor--of the order of a third of a standard deviation or less.

A preliminary, and rather quick literature search has turned up little more than discussion, and a few demonstrations, mostly on small samples. In one study, Ernest Lewis (1973) reported significant rises on IQ tests for 860 grade 6 students, more particularly on Verbal IQ, but without indicating size of effect. Welch and Walberg (1970) found no significant effects when 2,200 students from secondary schools were tested on such things as physics achievement, understanding science, process knowledge, etc. These students had an average IQ of 116. Lucas (1972) considered that their conclusions could be misleading, and found significant effects of up to one standard deviation using three samples of grade 12 students, each of about 50, and using the Watson-Glaser Critical Thinking Appraisal. Callenbach (1973), using 48 second grade students, found significant effects; he says "Although measurement experts...popular writers...and researchers...have described and analyzed test-wiseness (TW) and the effects of instruction in TW upon test performance, little of the TW literature and research has focused on the primary grades where, according to Joslin and others there has been a growing dependency upon group administered standardized tests...".

One interesting experiment is reported by Verster (1974) although its findings have limited applicability to our problems. A sample of 2,347 adults in an industrial setting with little test experience were given a battery of cognitive tests of ability. The group was then randomly divided into four subgroups. Group A were tested with the same battery four more times at three-month intervals; Group B were tested after an interval of six months, and then twice more at three-month intervals; Group C were tested after a nine-month interval, and again at the end of a year; and Group D had their first and only

posttest at the end of the year. The graphs below are typical learning curves.

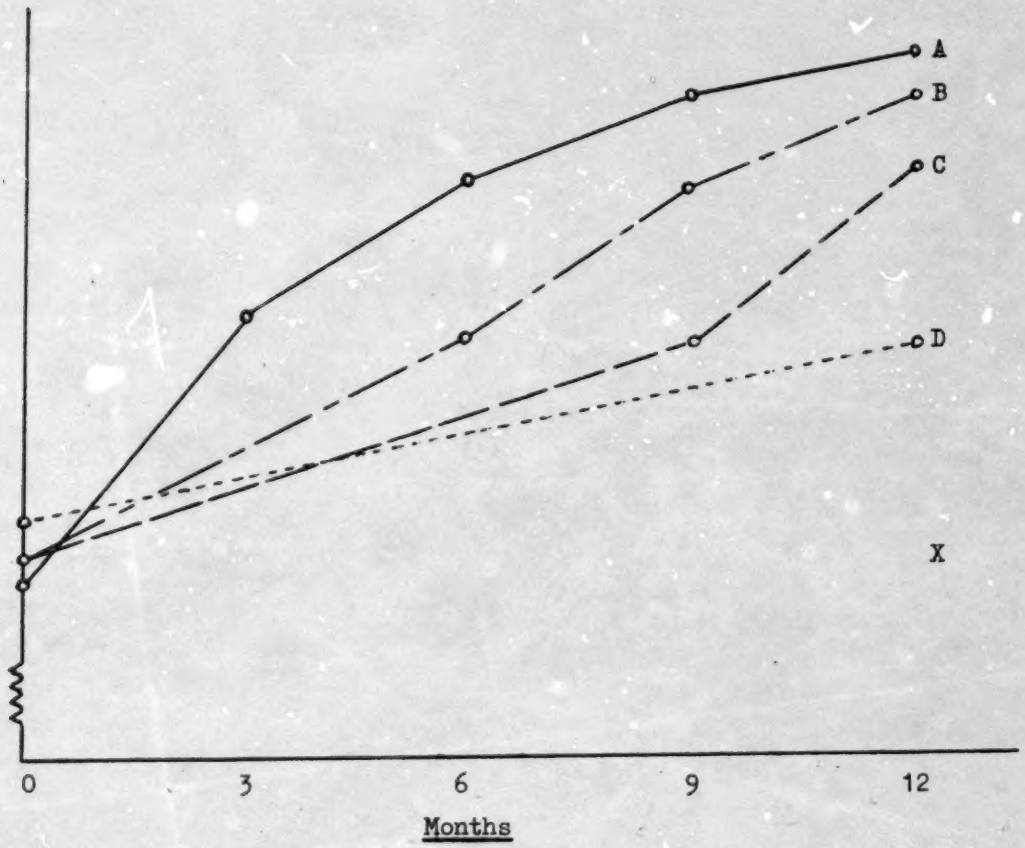

Figure 1. Example of Practice Effects

Notice that the first retest for all four groups produced almost identical gains (DX) irrespective of the time interval. This gain was roughly one third of a standard deviation; and even a lapse of a year had very little effect. The gain for the second retest is represented by the distance CD, and is again virtually constant for the three groups involved; it is about a quarter of a standard deviation. The gains BC and AB are smaller for third and fourth retests respectively, but for this sample, the total gain as the result of four retests is about one whole standard deviation. Even if one were to regard the first increase, DX, as being due to learning, the gain AD, about two thirds of a standard deviation, must be attributed to increase in test-taking skills alone.

If practice effect alone can cause important changes, then where gains for a treatment group only are being studied, some part at least would have to be discounted. The effect would be largest in the lower grades where there had been little test-taking; it would also be largest for less sophisticated students from countries with less emphasis on testing, for example perhaps, Portuguese immigrants from the Azores or Spanish-speaking Puerto Ricans.

Even where there are comparison groups, trouble can arise. For example, one program tested its students with two versions of the test for pretesting; for the posttest again two versions were given. Furthermore, the same students repeated this process in each subsequent grade so that by grade 4 they could have been tested more than sixteen times. The comparison group, however, was drawn randomly anew each year with, in all probability, a good deal lower average number of testings; at the least it received only one version of each test.

In another large program, a change occurred after the first series of tests. It was decided to equalize the effects of test sophistication for all its program students, on the grounds that some of them had had less exposure than others. They therefore called their teachers together, gave them a thorough briefing on the tests to be used later, instructed them to draw up their own parallel forms and to use these with standard instructions to give their pupils practice in doing these tests. Of course, their argument is correct that sufficient repetition would lift all students onto the plateau; but in converting retest scores to grade equivalents they are ignoring the fact that norms for that test were certainly not derived from such a well-trained population, so that gains from one year to the next were doubly spurious.

Lastly, it should not be thought that consideration of practice effect can only detract from positive findings; if the control group has had greater exposure to tests (as sometimes happens for precisely the same reasons that educational compensation was sought in the first place) then a straightforward comparison of mean test scores may underestimate the treatment effects.

Effects of Revisions of Test Norms

We have encountered this problem in more than one study. One program evaluator drew attention to the use of different standards in his report, but unless this is done it can easily escape being noticed.

Over the years, test publishers have sometimes found it necessary to

revise their norm tables. This has recently happened to the Stanford Achievement Tests amongst others. There appears to be a substantial lowering of standards; the same raw score now qualifies for higher grade equivalents, more particularly at the upper grades where the difference can be as much as a full grade or more higher than on the older norms. Modu and Stern (1975) found changes in the Scholastic Aptitude Tests of about a third of a standard deviation over the years 1963-1973. Whatever the reason for this, the use of the older norms at first testing or in lower grades, followed by conversions or new norms at retesting or in higher grades, can make the program appear to be a colossal success. When the tests themselves have been revised the same drop in standards undoubtedly exists, but is then even more difficult to detect or to compensate for.

"Post hoc, ergo propter hoc"

Ideally, of course, benefits from programs should be attributable solely to the effects of use of a selected theme, and increased expenditures should be warranted by this additional theme alone.

However, there can be little doubt that often the additional funding has been used for considerable improvement in conditions for the selected group only, and making it a moot point whether the use of a selected theme made any positive contribution at all. It is even conceivable that these improved conditions could have cloaked deterioration of performance in the select group, and could have produced more general benefits without the new element. It is not uncommon to find classrooms in a school with a teacher and an aide to cope with about 20 pupils, with tape recorders, slide projectors, Language Masters and shelves of new books, while next door a single teacher in a very ordinary classroom deals with 30 or more pupils. The problem these pupils in the ordinary classroom now have is that they did not have a problem to start with.

Of course it is seldom possible to identify the particular components of a treatment which have led to success; but it is a poor demonstration, when features well outside of the main theme are incorporated into the treatment, but kept away from the comparison group. Here, given the same additional facilities, the comparison group could conceivably outstrip the experimental group in performance.

An error that pops up now and then, in spite of repeated warnings, is that of inferring causal relationships wherever association is found. For example volunteer students are taught a foreign language, and it is found that

their average achievement in the home language is higher than that from non-volunteering students, from which it is "concluded" that learning a foreign language benefits performance in the home language, overlooking several more likely and more plausible explanations. Volunteers are seldom typical individuals.

Constraints on Implementation and Evaluation

Conflicts with Educational Ideals

The majority of evaluations in most studies seem to be flawed beyond redemption. It might be supposed, especially in view of the criticisms we have made of both experimental designs and of the statistical procedures used, that the evaluators lacked competence. It is, of course, true that the experimental design was usually that of the educators involved, and the statistical procedures were often selected by them. It is also true that research is a specialist occupation, and that there is no good reason whatever why teachers and educational administrators should be trained and experienced in both fields. But it would be wrong to assume that problems of evaluation would largely disappear with better training or more use of research specialists. Not only are many difficulties unavoidable, but it should be clearly recognized that modern educational practice and aims are often in direct opposition to the needs of sound research; that research is possible only by seeking and capitalizing upon observed differences, while modern education ideologically, if not ideally, considers differences a call to immediate action.

For example, while random allocation of a sample to experimental and control groups is a powerful statistical device, it is virtually impossible in most educational situations; on the contrary, placement in the treatment group is done precisely because there is a need to eliminate a difference between groups. Under threat even of court action, it is difficult to consider withholding cases for comparison, since if the treatment was effective, this would produce differences instead of removing them.

Conflicts with Laws and Regulations

Most education seeks to minimize differences between performances of individuals by raising the lower. There is, of course, an alternative philosophy; the goal could be, for each individual, to maximize the differences between his consecutive performances. However, we have not often encountered such an objective; almost all have sought to identify and treat groups whose performance

is below average. In at least one case, this purpose came into conflict with what the court considered to be the objectives of desegregation. Thus, not all failures must be laid at the door of poor design. The following case was not unique, and is an example of problems a program may have to face from federal, state, and court jurisdictions.

Their experimental design was as good as normal educational restraints permitted with commendable use of refined statistical procedures following, with frank interpretation of results, and with less special pleading than is common in this area. They included a control group in their design; checked on the initial comparability of control and experimental groups; recorded differences in exposure to their treatment; showed the effects on various subjects separately. They stated their hypotheses before analyzing their results. They avoided the pitfalls of multiple t-tests by using a more elaborate analysis of variance first. In the end, evidence for the success of their program was not overwhelming, but patently honest and entirely credible.

But then they were subjected to a series of interventions beyond their control. Over a five year period, they had at least six regional consultants with concomitant variations in interpretation of guidelines. A change in guidelines forced them to change from a planned horizontal expansion of their program to vertical expansion; they lost their control groups; they changed patterns of bussing, but found reduced contact between the two main groups of students. Next, they lost a desegregation suit to the Office of Civil Rights which forced them to close one school, to redistribute the students for whom the special treatment had been devised, and to reassign teaching staff.

They tried to readjust but then were subjected to a drastic cut in staff. They compensated by placing more emphasis on development of materials and in-service training, producing 27 specially trained teachers, all but seven of them paid from local funds--and lost 16 of them to wealthier districts when state legislation was enacted forcing the spread of the innovation. To cap it all, they then received a mandate to expand their program from grade 4 to grade 6.

This is not the only case in which special classes were judged to be in violation of desegregation guidelines. The effect of these decisions is obvious. Either the main thrust of the program must be considerably blunted, or else all students must be compelled to follow the same program, even those that do not need or want it. A court decision can alter the very nature of a program.

Curriculum Overload

Some innovational programs add appreciably to the work load of teachers and students. If the ordinary curricula already fill the time available, something will have to give. In only one case did we find evaluators vigilant enough to check progress in other subject areas. They found that less than one half of the set curricula in each of science and social studies had been completed in the year, and quite frankly attributed this to the effects of the increased work load.

Uncontrollable Sampling Biases

Ethical considerations if not indeed legal ones, force two limitations upon educationalists: They cannot easily deny students access to a program which is manifestly intended to confer some educational advantage; and they cannot easily override parental preferences even when they believe them misguided. Educationally this is perhaps often of not much consequence since in the long run many alternative systems lead to alternative goals of equal merit. It is quite a different matter when scientific demonstration hinges upon such decisions; then several undesirable interferences are probable, including significant sampling biases.

For example parents of some pupils will press, with a variety of motivations, for inclusion of their offspring in programs. Even when these pupils' results are considered separately, the constitution of comparison groups is almost certain to be compromised and in a way which will make the program appear better than it is. On occasion we suspected that the results had not been partitioned, and that would enhance the program's showing still further. Whatever the funding intent, some programs had considerable proportions of non-target pupils, sometimes as a result of active encouragement by the innovators. Even with target pupils, biases can, and demonstrably do occur as will be shown.

Two opposing considerations affect decisions by volunteering parents. Some parents seem understandably anxious not to interfere with a satisfactory progression through school, by changing horses in midstream. Thus especially at the start of a new program, those whose children have already acquired some skills prefer not to switch, while parents of children who are worried by a general lack of achievement see the new program as a new hope. This creates sampling differences which show results to the detriment of the program.

On the other hand when the program is firmly established and probably with an enriched environment and increased staff, the parents of the higher achievers seize the opportunity to transfer their children into the program and thus have the new experience as well. This kind of sampling bias produces data which show the program to advantage. We have encountered both trends in a single program at different stages of its development.

Conclusions

From the viewpoint of the research analyst, there seems to be room for improvement in communication between the administrative bodies of the funding offices, the school administrations, and the evaluation agencies. Nothing is going to produce substantial numbers of successful innovations; at this stage of educational development we can reasonably hope for few only, and only modest advances. However it would be a real advance if a substantial reduction could be made in the number of programs being rejected for <u>lack</u> of evidence, even if this meant an increase in the number disqualified by contrary evidence; this would at least increase the number to which serious consideration could be given.

References

Bowers, J. E., Campeau, P. L., & Roberts, A. O. H. Identifying, validating, and multi-media packaging of successful reading programs. Final report. Palo Alto, California: American Institutes for Research, December 1974. (AIR-41200-12174-FR).

Callenbach, C. The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. Journal of Educational Measurement, Spring 1973, 10 (1), 25-29.

Gage, N. Handbook of research on teaching. Chicago: Rand McNally, 1963.

Hawkridge, D. G., Chalupsky, A. B., & Roberts, A. O. H. A study of selected exemplary programs for the education of disadvantaged children, Parts I and II. Palo Alto, California: American Institutes for Research, September 1968.

Huck, S. W., & McLean, R. A. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. Psychological Bulletin, July 1975, 82 (4), 511-518.

Huff, D. Score: The strategy of taking tests. New York: Appleton-Century-Crofts, 1971.

Lewis, E. Repeated testing: Interpreting the results. Proceedings of the AERA Annual Meeting, Washington 1973, (summary).

Lucas, A. M. Inflated posttest scores seven months after pretest. Science Education, 1972, 56 (3), 381-387.

Millman, J., Bishop, C. H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25 (3), 707-726.

Modu, C. C., & Stern, J. The stability of the SAT score scale. Research Bulletin, College Entrance Examination Board Research and Development Reports, RB-75-9. Princeton, N. J.: Educational Testing Service, April 1975.

Porter, A. C., & Chibucos, T. R. Common problems of design and analysis in evaluative research. Sociological Methods and Research, Feb. 1975, 3 (3), 235-257.

Tallmadge, G. K. The development of project information packages for effective approaches in compensatory education. Los Altos, California: RMC Research Corporation, October 1974 (Technical Report No. UR-254).

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects (revised). Los Altos, California: RMC Research Corporation, December 1974.

Verster, M. A. The effects of mining experience and multiple test exposure on performance on the Classification Test Battery. Confidential report. C/PERS 220. Johannesburg, CSIR, NIPR, 1974.

Welch, W. W., & Walberg, H. J. Pretest and sensitization effects in curriculum evaluation. American Educational Research Journal, November 1970, 7 (4), 605-614.